NET 456 – High Speed Networks

Lecture 04
Networks in Switch Fabrics

Dr. Anis Koubaa
Outline

- Characteristics and Features of Switch Fabrics
- Crossbar Switch Fabrics
- Blocking Switch Fabrics
- Nonblocking Switch Fabrics: Clos Networks
- Concentration and Expansion Switches
- Shared-Memory Switch Fabrics
- Techniques for Improving Performance
Characteristics of Switch Fabrics
Characteristics of Switch Fabrics

- The switching function takes place in the switching fabric.
- The switching structure depends on the need of network operation, available technology, and the required capacity.
- A switch fabric is an interconnected network of smaller switching units.
- Several factors can be used to characterize switching systems:
  - buffering, complexity, capability of multipoint connections, speed, performance, cost, reliability, fault tolerance, and scalability.
Classification of Switching Networks

- The key factors for classifying switching networks are
  - **Single path vs. multipath**
    - A single-path network has exactly one route between each input port and output port
    - In a multipath network, any connection can be established through more than one path
  - **Fixed inter-stages vs. variable inter-stages**
    - In a single-stage network, a connection is established through one stage of switching
    - In a multistage network, a packet must pass through several switching stages
Classification of Switching Networks

- The key factors for classifying switching networks are
  - **Deflection routing vs. packet discard**
    - Deflected routing consists in routing packets using alternative output ports (paths) when a given output port is overloaded.
    - Packet discard means that packets are simply discarded in case of congested output port.
  - **Blocking vs. Nonblocking**
    - A switching network is said to be blocking if an input port cannot be connected to an unused output port. Otherwise, the switching network is nonblocking.
Classification of Nonblocking Switches

- A nonblocking switching network can be either of two types.
  - A network is **wide-sense nonblocking** if any input port can be connected to any unused output port without requiring a path to be rerouted.
  - A network is **rearrangeably nonblocking** if, for connecting an input port to an unused output port, rearranging other paths may be required.
Crossbar Switch Fabrics
Crossbar Switch Fabrics

- Crossbar switches are the building blocks of switching fabrics.
- In a crossbar, every input can be uniquely connected to every output through a **crosspoint**.
- A crosspoint is the smallest unit of the switching function.
  - can be built using a variety of electronic or digital elements, such as photodiodes, transistors, and AND gates.
- Crossbar switches are considered strictly (wide-sense) **nonblocking**, as a dedicated crosspoint exists for every connection.
  - The blocking, however, may occur when multiple packets are sent to the same output simultaneously.
Crossbar Switch Fabrics

- If a particular output port of the crossbar is idle, the desired connection can always be established by selecting the particular crosspoint dedicated to the particular input/output pair.

A crossbar switching fabric with $n = 4$ inputs/outputs
Crossbar Switch Fabrics

100-point, 6-wire crossbar switch. Manufactured by Western Electric April, 1970, near the end of the Type B production run. Similar switches were used in Incoming Register Links and other subsystems not part of the regular voice fabric. Note the Vertical Off Normal contacts above each hold magnet. The Horizontal Off Normal contacts were housed under transparent covers, to the left of the left butterfly magnets and to the right of the right butterflies.


Detail of back side of a Western Electric 6-wire 100-point crossbar switch (model 324-N) showing "banjo" wiring. Note this model of crossbar switch was divided into three groups of verticals; 4, 2, and 4. The banjo wiring as seen here is continuous only within each vertical group. More typically crossbar banjo wiring extended across all verticals of the switch.

Blocking Switch Fabrics
It is also possible to construct other multistage switch fabrics, with the trade-off of cost versus higher blocking.

Different types of blocking switches:

- Omega networks,
- Banyan networks,
- Delta networks,
- Bene networks.
In an Omega network, only one route exists between each input $i_x$ and output $o_x$.

The path uniqueness in this network can be shown by the induction on the number of stages, $s$.

- The number of stages is exactly $s = \log_d n$.

The Omega network has some useful unique properties:

- the interconnection segment between each two consecutive stages is identical to that of all other segments.
- Each of these identical interconnection segments is configured with a shuffling property. (cyclic logical left shift)
- The Omega network is also self-routing. The routing in this network is very simple and is achieved by comparing bit by bit between a source address and a destination address.
Routing in Omega Network

- The routing is achieved by comparing bit by bit between a source address and a destination address.
- At each stage, we look at the corresponding bits (starting at the most-significant bit (MSB)) of the source and the destination addresses.
  - If bits are the same, a message passes through
  - otherwise, it is crossed over.

(a) An Omega network with an example of its routing:

For example, consider a case with
- the source address 010
- the destination address 110

In the first stage, the message is crossed over, as the MSBs of the two addresses are different (0-1).
In the second stage, the next corresponding bits are the same (1-1), so the message is sent straight along the crossbar.
The case for the third stage is also the same as for the second stage but with corresponding bits of 0-0.
Banyan Network

- In electronics, a banyan switch is a complex crossover switch used in electrical or optical switches.
- Similar to Omega Network
- Slef Routing Capability

(b) a Banyan network with an example of its routing
Delta Network

- The Delta network consists of a network of crossbar switch elements with shared crosspoints, thus raising the possibility of blocking.
- In a Delta network, only one route between each input $i_x$ and output $o_x$ exists.

(a) A D8,2 Delta network with the realization of its routing;
(b) extension of the Delta network to a B8,2 Beneš
Non Blocking Switch Farbics
In telecommunications, a Clos network is a multistage circuit switching network, first formalized by Charles Clos in 1953, which represents a theoretical idealization of practical multi-stage telephone switching systems.

Clos networks are required when the physical circuit switching needs exceed the capacity of the largest feasible single crossbar switch.

The key advantage of Clos networks is that the number of crosspoints (which make up each crossbar switch) required can be much fewer than were the entire switching system implemented with one large crossbar switch.

NET 456: High Speed Networks, by Dr. Anis Koubaa
Clos Networks (source: Wikipedia)

- Clos networks have three stages:
  - the ingress stage,
  - the middle stage,
  - the egress stage.

- Each stage is made up of a number of crossbar switches often just called crossbars.

- Each call entering an ingress crossbar switch can be routed through any of the available middle stage crossbar switches, to the relevant egress crossbar switch.

- A middle stage crossbar is available for a particular new call if both the link connecting the ingress switch to the middle stage switch, and the link connecting the middle stage switch to the egress switch, are free.
Clos networks are defined by three integers $n$, $m$, and $r$.
- $r$ number of ingress stage of crossbar switches
- $n$ represents the number of sources which feed into each of $r$ ingress stage crossbar switches.
- $m$ number of outlets in each ingress stage crossbar.
Features of Clos Networks (source: Wikipedia)

- There is exactly one connection between each ingress stage switch and each middle stage switch.
- There are $r$ egress stage switches, each with $m$ inputs and $n$ outputs.
- Each middle stage switch is connected exactly once to each egress stage switch. Thus, the ingress stage has $r$ switches, each of which has $n$ inputs and $m$ outputs.
- The middle stage has $m$ switches, each of which has $r$ inputs and $r$ outputs.
- The egress stage has $r$ switches, each of which has $m$ inputs and $n$ outputs.
Clos Networks

A Clos network with
• $r = 4$ ingress stages
• $n = 2$ input per each ingress stage
• $m = 5$ outlets for each ingress stage
Clos Networks

_Clos network for 128 hosts with all fiber ports and monitoring capability_

This family of modular, Myrinet-2000, switch products is recommended for clusters with up to 128 hosts. If your cluster is likely to grow to more than 128 hosts, see also the switch-network products in the following section, which provide higher port density and enterprise features that are not available in this original series of Myrinet-2000 switches.

Clos Networks (source: Wikipedia)

- The relative values of $m$ and $n$ define the blocking characteristics of the Clos network.
  - **Strict-sense nonblocking Clos networks ($m \geq 2n - 1$)** (Clos's classic 1953 paper)
    - If $m \geq 2n - 1$, the Clos network is *strict-sense nonblocking*, meaning that an unused input on an ingress switch can always be connected to an unused output on an egress switch, *without having to re-arrange existing calls*.
  - **Rearrangeably nonblocking Clos networks ($m \geq n$)**
    - If $m \geq n$, the Clos network is *rearrangeably nonblocking*, meaning that an unused input on an ingress switch can always be connected to an unused output on an egress switch, but for this to take place, existing calls may have to be rearranged by assigning them to different centre stage switches in the Clos network.
Exercise: Blocking Probability

- Let us consider a Clos Network where $m = n$. The objective of this exercise is to calculate the blocking probability, which is the probability that there is no available link between an ingress switch and egress switch.

- We assume that
  - the number of other active calls on each ingress or egress switch is $u = n - 1$.
  - there is no possible rearrangement of calls.
  - $p$ is the probability that each internal link between stages is already occupied by a call. This probability is completely independent between different links. (Lee Approximation).
  - $q$ is the probability that an ingress or egress link is busy.
Exercise: Blocking Probability

- Express the probability $p$ as a function of $q, m$ and $u$.
- What is the probability that the path connecting an ingress switch to an egress switch via a particular middle stage switch is free?
- What is the probability of blocking in the switch?
Concentration and Expansion Switches
Concentration and Expansion Switches

- Some switching networks can either expand or concentrate incoming traffic.
  - Concentration Switches
  - Expansion Switches

- In a **concentration-based switching network**, the number of output ports is less than the number of input ports.

- In an **expansion-based switching network**, the number of output ports is more than the number of input ports.
Knockout Switch

- One of the simplest concentration switching networks
- The knockout switch is a blocking network
- It is constructed with interconnected crossbars with less complexity than a crossbar switch of the same size.
- The knockout switch is a good substitution for a crossbar switch when the likelihood is small that many inputs will need to send packets to the same output simultaneously.

The knockout switching network
Knockout Switch

- The idea of this switch is based on the concept of knocking a possible \( k \) packets out of \( n \) active inputs if the number of switch outputs is \( m \), where \( m < n \).
- The switch elements act as concentrators, ensuring fairness to all entering packets.

The knockout switching network

\[ n = 8 \text{ inputs} \]
\[ m = 4 \text{ outputs} \]
Knockout Switch

- The eight incoming packets from eight switch inputs compete in the first four switch elements.
- A packet that wins the competition, based on a random selection in a switch element, stays in the same column of switches and continues to reach its desired output.
- Otherwise, a losing packet is knocked out to the next column of switch elements for the next round of competition.
- The first-round losers in this process go straight to the second column and play off against one another and at later stages, face the second-and the third-round losing packets from the first column.
Expansion Network

- An expansion network with \( n_1 \) inputs and \( n_2 \) outputs, where \( n_1 < n_2 \), can scale up the number of receivers from \( n_1 \) to \( n_2 \).
- An expansion network is a generalization of the three-stage Clos network and is useful in applications in which an incoming signal is to be sent to multiple outputs.
- Such networks are also called distribution networks.

A three-stage expansion switching network
Shared-Memory Switch Fabrics
Shared-Memory Switch Fabrics

- A totally different approach for switching can be achieved in time domain and without using any switch elements: the shared-memory switch fabric.
- The distinctive feature of this switch fabric is the use of a high-speed $n:1$, time-division multiplexer (TDM) with a bit rate $n$ times as large as the rate on each individual input/output line.
Figure 13.11

Shared-memory switch fabric